

Πρώτη Σειρά Ασκήσεων – Μέρος Α

Σε αυτή την άσκηση θα πρέπει να τροποποιήσετε τον αλγόριθμο Reservoir Sampling που περιγράψαμε στο μάθημα, ώστε να κάνει δειγματοληψία K αντικειμένων ομοιόμορφα τυχαία από ένα ρεύμα N αντικειμένων. Το κάθε αντικείμενο πρέπει να έχει πιθανότητα K/N να εμφανιστεί στο δείγμα. Ο αλγόριθμος σας θα πρέπει να δουλεύει με ένα μόνο πέρασμα στα δεδομένα διαβάζοντας τα αντικείμενα ένα-ένα, χωρίς προηγούμενη γνώση του μεγέθους του ρεύματος (το μέγεθος N), και να χρησιμοποιεί $O(K)$ μνήμη (υποθέστε ότι το μέγεθος του κάθε αντικειμένου είναι σταθερό). Αυτό σημαίνει ότι δεν μπορείτε να αποθηκεύσετε όλο το ρεύμα των δεδομένων στη μνήμη.

1. Περιγράψετε τον αλγόριθμο που διαλέγει ένα ομοιόμορφο δείγμα K αντικειμένων από ένα ρεύμα N αντικειμένων. Η περιγραφή του αλγορίθμου δεν πρέπει να είναι σε κώδικα ή ψευδοκώδικα, ούτε η περιγραφή του κώδικα σε φυσική γλώσσα. Στην περίπτωση αυτή η απάντησή σας μηδενίζεται. Η περιγραφή θα πρέπει να εξηγεί τη λογική του αλγορίθμου σε απλά Ελληνικά.

Για παράδειγμα, αυτή είναι μία περιγραφή σε απλά Ελληνικά του αλγορίθμου για τη δειγματοληψία ενός αντικειμένου που είδαμε στην τάξη:

Ο αλγόριθμος κρατάει χώρο για ένα αντικείμενο και ένα μετρητή με τον αριθμό των αντικειμένων που έχει δει. Διατρέχει τα αντικείμενα ένα-ένα όπως έρχονται από το ρεύμα. Όταν βλέπει το n -οστό αντικείμενο, το επιλέγει με πιθανότητα $\frac{1}{n}$ και το αποθηκεύει, αντικαθιστώντας το υπάρχον αντικείμενο (αν δεν είναι το πρώτο). Ενημερώνει τον μετρητή. Όταν ολοκληρωθεί το ρεύμα, επιστρέφει το αντικείμενο που έχει αποθηκεύσει.

2. Αποδείξτε ότι ο αλγόριθμος σας παράγει ένα ομοιόμορφα τυχαίο δείγμα, δηλαδή, για κάθε $i, 1 \leq i \leq N$, το i -οστό στοιχείο έχει πιθανότητα K/N να εμφανιστεί στο δείγμα.
3. Γράψτε μία συνάρτηση **sample** σε **Python** που υλοποιεί τον αλγόριθμο σας. Η συνάρτησή σας θα πρέπει να παίρνει σαν όρισμα το όνομα ενός αρχείου και τον αριθμό K , και να επιστρέφει μια λίστα που κρατάει ένα δείγμα με K τυχαίες γραμμές από το αρχείο. Θα πρέπει να διαβάσετε το αρχείο γραμμή-γραμμή και να μην το φορτώσετε στη μνήμη. Χρησιμοποιήστε την συνάρτησή σας μέσα σε ένα πρόγραμμα για να πάρετε 10 τυχαίες γραμμές από το αρχείο *input-A.csv* που θα σας δίνεται. Εκτυπώστε τις γραμμές στο δείγμα.

Εξηγήστε την αντιστοίχιση μεταξύ της περιγραφής που δώσατε στο 1^ο βήμα και του κώδικα σας. Για παράδειγμα, για τον αλγόριθμο που επιλέγει ένα μόνο αντικείμενο, θα μπορούσατε να γράψετε κάτι της μορφής: «Στις γραμμές 2-3 γίνεται η επιλογή του n -οστού αντικειμένου με πιθανότητα $\frac{1}{n}$ ».

Δημιουργήστε ένα Notebook με δύο κελιά κώδικα. Ένα με τα imports και τον ορισμό της συνάρτησης **sample** και ένα στο οποίο θα χρησιμοποιείτε την συνάρτησή σας στο αρχείο *input-A.csv*, θα αποθηκεύετε το αποτέλεσμα και θα το εκτυπώνετε. Μπορείτε να κατεβάσετε το αρχείο *input-A.csv* από την σελίδα Ασκήσεις. Προσθέστε και δύο κελιά κειμένου, ένα με την περιγραφή του αλγορίθμου (Βήμα 1), και ένα με την

αντιστοίχιση μεταξύ κώδικά και περιγραφής. Μπορείτε να προσθέσετε την απόδειξη (Βήμα 2) σε ένα ξεχωριστό κελί, ή να την γράψετε ξεχωριστά και να παραδώσετε ένα pdf με το κείμενο (ή φωτογραφίες αν είναι χειρόγραφο). Παραδώστε το Notebook, και το pdf (ή φωτογραφίες) με την απόδειξη αν υπάρχει.